К ПРОБЛЕМЕ ОЦЕНКИ КАЧЕСТВА ПЕДАГОГИЧЕСКОГО ТЕСТА

Инна Щербинина
Морской государственный университет
им. адм. Г.И. Невельского, г. Владивосток
inna@mail.ru

Статья посвящена описанию метода анализа соответствий и особенностей его применения для оценки качества педагогического теста по результатам пробных испытаний.

Ключевые слова: *надёжность теста, дифференцирующая способность тестового задания, методы оценки качества педагогического теста.*

Проблемам оценки качества теста, его надёжности и дифференцирующей способности посвящено немало исследований. Для получения численной оценки этих качеств теста выделяются два подхода:

- исследование только результатов тестирования, не сравнивая их с результатами других форм педагогического контроля;
- сравнение результатов тестирования с результатами применения какой-либо другой формы педагогического контроля.

В рамках первого подхода используются такие широко распространённые методы, как вычисление так называемого point-biserial коэффициента корреляции, характеризующего дифференцирующую способность тестового задания:

$$\tau_{pbis} = \frac{(\bar{x}_1)_j - (\bar{x}_0)_j}{Sx} \cdot \sqrt{\frac{N_1 \cdot N_0}{N(N-1)}}, \qquad (1)$$

где, $(\bar{x}_1)_j$ — средний балл студентов, выполнивших j-е задание;

$(\bar{x}_0)_j$ — средний балл студентов, не выполнивших j-е задание;

1

Описание метода и пример использования можно найти в работе Панченко А.А. Разработка тестов // Хабаровск: 2000.

Sx — среднеквадратическое (стандартное) отклонение баллов по всей выборке;

N_1 — количество студентов, верно выполнивших j-е задание;

N_0 — количество студентов, неверно выполнивших j-е задание;

N — общий объём выборки[1].

Оценка надёжности дихотомических заданий путём применения формулы Кьюдера–Ричардсон (KR-20), основанной на статистической оценке согласованности ответов по всем заданиям теста:

$$\tau_{_н} = \frac{n}{n-1} \cdot \left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{Sx^2} \right), \quad (2)$$

где, n — число заданий теста;

p_j — доля правильных ответов на j задание теста;

q_j — доля неправильных ответов на j задание теста;

Sx^2 — дисперсия баллов по всему тесту.

В некоторых работах указывается возможность использования коэффициента корреляции при делении множества испытуемых на две подгруппы или такого же деления множества тестовых заданий (например, на чётные и нечётные)[2].

Второй подход основан на применении коэффициента корреляции. О результатах, получаемых в случае применения второго подхода, и их интерпретации хотелось бы поговорить особо.

Для сравнения отметок, полученных путём применения различных форм педагогического контроля обычно используется коэффициент корреляции, характеризующий наличие и степень зависимости между сравниваемыми числовыми последовательностями. Использование для оценки знаний чисел 2, 3, 4, 5, называемых отметками, достаточно условно. Недаром в высшей школе для тех же целей используются термины «неудовлетворительно», «удовлетворительно», «хорошо», «отлично». Отметки позволяют сравнивать знания обучающихся, но для этих чисел не определено сложение и вычитание, равно как умножение и деление, поскольку две «двойки», полученные по предмету, не равны «четвёрке». Это объясняется тем, что измерения проводятся в порядковой шкале. Фактически, это оценка знаний, которая позволяет утверждать, что отличник знает предмет лучше, чем хорошист, но не позволяет оценить, насколько. Для сравнения данных, представленных в порядковых шкалах, допустимо использование коэффициент корреляции Спирмена, или коэффициента корреляции Кендела. В педагогических измерениях использование коэффи-

2

Майоров А.Н. Тесты школьных достижений: конструирование, проведение использование // С-Пб.: Образование и культура. 1996.

циента корреляции Спирмена предпочтительнее, поскольку он более чувствителен к небольшим отклонениям.

Для определения статистической значимости полученного коэффициента корреляции считают величину

$$t_{рас} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}},$$

где, r — коэффициент корреляции;

N — мощность выборки.

Для подтверждения значимости коэффициента корреляции необходимо сравнить $t_{рас}$ и $t_{таб}$ (табличное значение распределения Стьюдента). Если расчётное значение $t_{рас}$ больше табличного, взятого с $N-2$ степенями свободы, то отвергается нулевая гипотеза, т.е. гипотеза о том, что данные последовательности независимы. Однако статистически значимый коэффициент корреляции не даёт гарантии надёжности теста, т.е. возможности получения устойчивой оценки способностей испытуемых.

Проиллюстрируем сказанное следующим примером. Мощность выборки, на которой производилось измерение, равно 90 человек. И результаты теста, и результаты экзамена представлены в виде выставленных традиционных отметок. Для интерпретации «сырых баллов» теста использовалась равномерная шкала, приведённая в табл. 1.

В табл. 2 приведены данные по сравнению результатов теста и экзамена. Тестирование проводилось непосредственно перед экзаменом. Отметка, полученная на экзамене, не зависела от набранных при тестировании баллов.

Табл. 2 является таблицей соответствия признаков и строится следующим образом: в столбцах таблицы приведены отметки экзамена, в строках — результаты тестирования, любая ячейка содержит количество человек, получивших указанную в столбце отметку на экзамене и указанную в данной строке отметку при тестирова-

Таблица 1. **Соответствие результатов теста традиционным отметкам**

Результаты теста	Традиционная отметка
0,00–0,40	неудовлетворительно (2)
0, 41–0,60	удовлетворительно (3)
0,61–0,80	хорошо (4)
0,81–1,00	отлично (5)

Таблица 2. **Распределение отметок экзамена в рамках отметок теста**

	Неуд.	Уд.	Хорошо	Отлично	Итого (тест)
Неуд.	26	6	4	0	36
Уд.	12	1	4	2	19
Хорошо	6	10	8	4	28
Отлично	0	1	3	3	7
Итого (экзамен)	44	18	19	9	90

нии. Например: 2 человека получили отметку «отлично» на экзамене и отметку «удовлетворительно» с помощью тестирования.

Коэффициент корреляции Спирмена рассчитывается по формуле:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}, \qquad (3)$$

где, r_s — значение рангового коэффициента корреляции Спирмена;

d_i — разность между значениями рангов в i-й строке;

n — объём выборки[3].

Коэффициент корреляции Спирмена равен 0,531 и является статистически значимым при доверительной вероятности 0,99 (для выборки с 88 степенями свободы).

Однако, рассматривая соответствие отметок теста и экзамена, можно заметить, что совпадение по неудовлетворительным отметкам достаточно велико (72% случаев), а совпадение по остальным отметкам менее значимо (для удовлетворительных отметок — 5%, для «хорошо» — 28,6%, для «отлично» — 42,9%). Хороший же показатель коэффициента корреляции обуславливается значительным пересечением неудовлетворительных отметок и их большим числом. По мнению автора, о надёжности теста следует говорить при наличии статистически значимой зависимости для всех отметок.

Для определения наличия зависимости по каждой из отметок может быть использован метод анализа соответствий, основанный на применении критерия χ^2.

Этот метод состоит в построении для каждой из отметок таблицы соответствия 2 × 2.

f_{11}	f_{12}	$f_{1\bullet}$
f_{21}	f_{22}	f_2
$f_{\bullet 1}$	$f_{\bullet 2}$	n

где, f_{11} — количество случаев присутствия обоих признаков (т.е. пересечения отметок разных форм педагогического контроля);

f_{12} — количество случаев наличия первого признака и отсутствия второго;

f_{21} — количество случаев отсутствия первого признака и наличия второго;

f_{22} — количество случаев отсутствия обоих признаков;

$f_{1\bullet}$ — сумма по первой строке таблицы;

$f_{2\bullet}$ — сумма по второй строке таблицы;

$f_{\bullet 1}$ — сумма по первому столбцу таблицы;

$f_{\bullet 2}$ — сумма по второму столбцу таблицы;

n — мощность выборки.

Для неудовлетворительных отметок рассматриваемого примера результаты расслоения приведены в табл. 3.

3

Гласс Дж. Стенли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976.

Таблица 3. **Данные для применения метода анализа соответствий неудовлетворительных отметок**

26	10	36
18	36	54
44	46	90

Экспериментальное значение рассчитывается по следующей формуле: [4]

$$\chi_0^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\bullet}f_{2\bullet}f_{\bullet 1}f_{\bullet 2}}. \quad (4)$$

Для данного примера

$$\chi_0^2 = \frac{90\cdot(26\cdot 36 - 10\cdot 18)^2}{36\cdot 54\cdot 44\cdot 46} = 13{,}07.$$

Полученное значение χ_0^2 сравниваем с критическим значением $\chi^2_{крит}$ для определённой вероятности ошибки первого рода α и одной степени свободы v, взятой из справочников по математической статистике. Так, $\chi^2_{крит}$ ($\alpha = 0{,}01$; $v = 1$) = 6,6. Если $\chi_0^2 < \chi^2_{крит}$, можно говорить о независимости получения одинаковых отметок с помощью исследуемых форм педагогического контроля, в противном случае подтверждается гипотеза о наличии зависимости полученных отметок.

Результаты исследования остальных отметок приведены в табл. 4.

Сравнивая с $\chi^2_{крит}$, видим, что статистически значимо пересечение только для отметок «неудовлетворительно» и «отлично».

На наш взгляд, данный тест даёт надёжные результаты только для неудовлетворительных и отличных отметок и поэтому может использоваться только как допуск к традиционному экзамену, освобождая преподавателя от необходимости тратить время на общение со студентами, по каким либо причинам неготовым к экзамену. При этом недопустимо его использование как альтернативы экзамену.

Этот же пример демонстрирует нечувствительность других, более общих методов.

Таблица 4. **Данные и результаты применения метода анализа соответствия для отметок «удовлетворительно», «хорошо» и «отлично»**

Удовлетворительно			Хорошо			Отлично		
1	18	19	8	20	28	3	4	7
17	54	71	11	51	62	6	77	83
18	72	90	19	71	90	9	81	90
$\chi_0^2 = 3{,}27$			$\chi_0^2 = 1{,}36$			$\chi_0^2 = 9{,}11$		

4

Афифи А., Эйзен С. Статитстический анализ // М.: Мир: 1982.

Так, обобщением метода анализа соответствий может служить метод χ^2 Пирсона таблиц сопряжённости признаков для $c \times r$ слоёв[5]. Данные для его применения приведены в табл. 2. Обозначения совпадают с приведёнными для таблиц 2×2.

Для каждой ячейки таблицы считается ожидаемая частота

$$F_{ij} = \frac{f_{i\bullet} f_{\bullet j}}{n}. \qquad (5)$$

А значение χ_0^2

$$\chi_0^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(f_{ij} - F_{ij})^2}{F_{ij}}. \qquad (6)$$

Полученное значение так же сравнивается с $\chi^2_{\text{крит}}$ для $\nu = (c-1)(r-1)$ степеней свободы. Для приведённого примера $\chi_0^2 = 34{,}6$. Данный результат показывает наличие зависимости между результатами компьютерного теста и традиционного экзамена. Но ранее мы показали, что этот результат значим только для отметок «неудовлетворительно» и «отлично».

О границах применения метода анализа соответствий. В литературе при исследовании этого метода для таблиц 2×2 указывается необходимость внесения поправки Йетса при $n \le 40$[6]. В этом случае формула (4) принимает вид:

$$\chi_0^2 = \frac{n \cdot (\left| f_{11} f_{22} - f_{12} f_{21} \right| - \frac{n}{2})^2}{f_{1\bullet} f_{2\bullet} f_{\bullet 1} f_{\bullet 2}}. \qquad (7)$$

Где $\left| f_{11} f_{22} - f_{12} f_{21} \right|$ — модуль разности.

При $n \le 20$, особенно в случае, когда хотя бы одно $f_{ij} < 5$, применение метода анализа соответствий к таблицам сопряжённости признаков 2×2 даёт невысокую точность. В этом случае рекомендуется применение точного критерия Фишера[7]:

$$\Pr = \frac{f_{1\bullet}! \cdot f_{2\bullet}! \cdot f_{\bullet 1}! \cdot f_{\bullet 2}!}{f_{11}! \cdot f_{12}! \cdot f_{21}! \cdot f_{22}! \cdot n!}. \qquad (8)$$

Данный критерий даёт точное значение ожидаемой частоты в ячейке, в то время как критерий χ^2 даёт приближённое значение.

5 *Афифи А., Эйзен С.,* Статистический анализ // М.: Мир: 1982.

6 *Кендал М. Стюарт А.* Статистические выводы и связи // М.: Наука: 1973.

7 *Г. Антон* Анализ таблиц сопряженности.